Facial Expression Synthesis Based on Natural Voice for Virtual Face-to-Face Communication with Machine

[†] Shigeo MORISHIMA and [‡] Hiroshi HARASHIMA
[†] Seikei University and [‡] University of Tokyo

[†] Faculty of Engineering, Seikei University
3-3-1 Kichijoji-kitamachi, Musashino Tokyo 180, Japan
Phone:+81-422-37-3726   Fax:+81-422-37-3871
E-mail:shigeo@tansei.cc.u-tokyo.ac.jp
[‡] Faculty of Engineering, University of Tokyo
7-3-1 Hongo Bunkyo-ku Tokyo 113, Japan

Abstract -- Basic research to a virtual face-to-face communication environment between an operator and a machine is presented. In this system, human natural face appears on the display of machine and can talk to operator with natural voice and natural face expressions. Especially in this paper, face expression synthesis scheme driven by natural voice is presented. Voice is including not only linguistic information but also emotional features. So we proposed expression control scheme driven by both features. We express a human head with 3D wire frame model. The surface model is generated by texture mapping with 2D real image. All the motions and expressions are synthesized and controlled automatically by the movement of some feature points on the model.

1. INTRODUCTION[1][2][3][4][5][12]

A user friendly human-machine interface using Multi-media are focused recently. Our goal is to realize very natural human-machine communication environment by giving a face to computer terminal or communication system. It is virtual face-to-face communication system between user and machine. For this purpose, we have already proposed basic schemes including a 3D modeling of face, face expression synthesis and coding technique, media conversion schemes, modeling and rendering method of hair, and modeling method of emotional aspect of facial expression based on the parameter mapping using neural network. A real-time animation synthesizer based on Pixel Machine is constructed for the interface prototype system. In these schemes, *Facial Action Coding System (FACS)* is selected as the efficient criteria to describe delicate face expression and motion.

Voice is essential to multi-media interface. It's including linguistic information, speaker information, emotional information and so on. If these information can be extracted automatically, natural voice can be an information source of human-machine communication. Moreover, media conversion and media integration of multi-media can be realized in the semantic level.

This paper presents a facial expression synthesis scheme driven by natural voice. Especially, two aspects of voice are utilized to express face image. Mouth shape and its motion are controlled by the linguistic information of voice. Full face expressions are controlled by the basic emotional information included in natural voice.

In this paper, some basic schemes to synthesize natural face expression, synchronization method between synthesized motion image and natural voice, scenario making tool to express delicate face expressions and animation scenes and quantitative emotion model based on neural network are presented. The basic research of emotion extraction from natural voice is reported at last.

0-7803-1363-1/93 $3.00 © 1993 IEEE

2. *EXPRESSION SYNTHESIS*[8][13]

2.1. Modeling of Human Face

A 3-D generic model which approximately represents a human face, is composed of about 600 polygonal elements and was constructed by measuring a mannequin's head. This generic model is including the teeth's model inside the head. The generic model is 3-D affine-transformed to harmonize its several feature point positions with those of given 2-D full-face image. This point adjustment is done by semi-automatical procedures. Some feature points' positions around face, lip, eyes and eyebrows are recognized roughly using the results of color information analysis of original image. Some corrections of each feature point can be done manually if necessary.

RGB intensity for a 2-D full-face surface image is then projected and mapped onto an adjusted generic model, following which a 3-D personal facial model is created. This model has a set of points which have 3-D oblique coordinate values and intensity in every polygon.

Once the 3-D facial model is gotten from the 2-D original image, it's easy to rotate the 3-D model in any arbitrary direction or to give many delicate facial actions for lips, jaws, eyes, and eyebrows by controlling lattice points in the wire frame.

2.2. Facial Expression Synthesis

Facial image synthesis is composed of deforming the wire frame model through facial expression parameters and mapping the texture in every polygon of the original image onto the surface of the deformed wire frame model. The deforming rules are formulated to simulate the facial muscular actions. Ekman and Friesen decomposed the facial muscular actions into 44 basic actions called *"Action Units" (AUs)* and described the facial expressions as combinations of *AUs* .

We express each *AU* with the combination of the movements of some specific feature points in the wire frame model. We quantize the movement in each AU into 100 levels between maximum change and no change. So any intermediate expression can be synthesized by combination of *AUs* and their levels.

2.3. High Definition Model

Now, we have a high definition wire frame model (HD model) which is measured by 3D digitizer (Cyberware) and has 24000 polygons and surface texture all around the head. Figure 1 shows both generic model (600 polygons version) and HD model. However, HD model doesn't have every feature point explicitly, so HD model has to be controlled hierarchically using generic model after adjustment between generic model and HD model. This adjustment is done by analysis of front view texture synthesized by HD model.

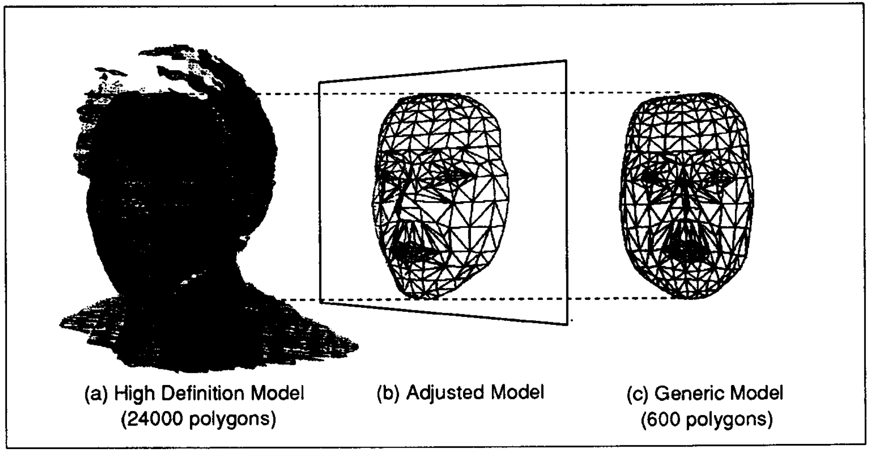

Figure 1. Hierarchical control of HD model with generic model

3. MEDIA CONVERSION SCHEMES[3][6][7][10][11]

When one is speaking, the motion around mouth can be predicted and synthesized only by text or voice. We have already proposed two types of media conversion schemes. They are Text-to-Image conversion and Voice-to-Image conversion. In the former system, mouth shape and motion are synthesized automatically by given Japanese or English text sentence. As the Voice-to-Image conversion system, we proposed two systems. These are frame by frame style conversion system based on neural network for real-time communication and recognition based conversion system for voice storage or voice mail. In the multi-media e-mail system, there may be both natual voice and mail text, so high quality mouth shape control and synchronization of voice and image can be achieved.

3.1. Text to Image Conversion

When each phoneme of sentences is analized to the *Allophone* code, a standard mouth shape and duration time are decided by the table of this allophone symbol. The number of typical mouth shapes are 17 categories in Japanese and 65 categories in English. In some consonants, the mouth shapes should be decided by an interpolation of the parameters between the preceding and following phonemes. Typical mouth shapes are located on the keyframes decided by the standard duration of each phoneme.

The motion between keyframes is decided upon by 3-D basis Spline interpolation of feature points. So, lip motions change very smoothly and naturally. In English, pronunciation is more complicated than Japanese and phonetic symbol and its duration time change according to word and accent. So, our dictionary much more categories for English.

To construct a Multi-media interface using this scheme, speech synthesis system is indispensable to make voice synchronize with synthesized motion image. Phonetic symbol and its duration are used at both image synthesis system and speech synthesis system.

3.2. Voice to Image Conversion for Communication System

For the voice communication system, real-time processing of speech and synthesis of image are essencial. Here, mouth shape is generated frame by frame basis with speech spectrum analysis. The window size is 32 msec. and frame rate is 1/30 second same as the video rate, so synchronization can be done simultaneously when there is no delay between video output and speech playback. Media conversion is progress based on three layer feedforward neural network. The input layer has 16 units, corresponding to the dimensions for LPC Cepstrum parameters. The output layer units correspond to lip control parameters. The mapping is learned by only a few training sequence, but the motion can be generated very delicately because the interpolation effect of neural network appears.

We have already construced a prototype voice communication system based on pipe line processing using Pixel Machine of AT&T. When one speaks in front of the microphone, the processes from speech analysis to image synthesis go at the speed of 10 frames per second with few delay time.

3.3. Voice to Image Conversion for Voice Storage System

In case of a media conversion from stored voice to image, the processing delay time isn't so important. So, complicated algorithms like speech segmentation and vowel recogintion can be available to improve the conversion performance. In this system, focused on the spectrum change and power shift between successive frames, segmentation boundaries are founded and keyframes are located based on this result. Mouth shape of each segment is decided by template matching of vowel. So only 8 categories are included in the templates. Interpolation of keyframes are 3D Spline function. Subjective evaluations of this synthesized motion image show higher conversion performance than the neural network version. When speech analysis and image synthesis can proceed in off-line condition, prototype system can be implemented on low-cost workstation .

3.4. Voice and Text to Image Conversion for Advanced Mail System

In this section, synchronization method of animation and natural voice using both stored voice and text is introduced. Voice recognition for long sentence is difficult problem now. So vowel recognition using template matching and voice sequence segmentation are performed as bottom-up process. After that process, text of sentence is entered and time scale matching between recognized vowels and entered vowels is performed. At last, consonant position is decided by entered text and standard duration table of phoneme. In this system, mouth shapes can be generated accurately by text information and duration of each phoneme can be decided by segmentation result. In the advanced multi-media e-mail system, text input and voice input may be possible.

3.4.1. Vowel Recognition and Segmentation

Vowels in the long Japanese sentence is detected here. Recognition categories are five Japanese vowels that are /a/, /i/, /u/, /e/, /o/, nasals /N/ and voiceless. At first, voice sequience is stored and devided into many segments according to power change and spectrum distance between successive two frames. Each segment is classified into a vowel category by template matching. This result includes more numbers of segments than real phoneme numbers in voice and it has some vowel recognition errors too.

3.4.2. Time Scale Matching

When text sentence is entered, the order and number of phoneme become clear. DP based time scale matching proceeds by comparing a recognized vowel sequence and a correct vowel sequence of input text. Correct vowel sequence is located on the vertical axis and recognized vowel sequence is located on horizontal axis. Matching path takes zigzagging line. The distance between each phoneme category is pre-determined as normalized Cepstrum distance using training data of specific speaker. The optimum path is decided to minimize the total distance between correct sequence and recognized sequence using DP Matching method. The vowel recognition result includes many errors, but these errors can be restored and more accurate segment position of each phoneme of text are decided by this matching.

Each consonant position is located in front of this vowel position and its length is selected by standard duration table. The keyframes for standard mouth shape are located at the start point of each phoneme segment. The shape of mouth between keyframes is interpolated by 3D Spline function. Mouth closing moments can be re-generated as well as original motion.

4. SCENARIO MAKING TOOL[13][14]

Delicate face expression and facial animation synthesis scenario can be generated by scenario making tool. This tool gives an animation making environment to user.

Facial expression and motion around mouth are controlled by several media conversion schemes selected. The other expression is controlled manually in this tool. The face expression can be chosen from the database and put it into time axis window by mouth operation. Standard expressions are stored as the combination of AU's numbers and its intensities. If necessary, user can preview and check the motion image on wire frame model.

After choice of original 2-D face image and execution of texture mapping processes, all frame motion pictures can appear on the window. The *Facial Action Coding System (FACS)* is a well-known method to describe facial actions. However, *FACS* doesn't provide a numeric representation. In this system, any kinds of face expressions can be realized by combination of *Action Units* of *FACS* and these intensities quantized by 100 levels.

This scenario making tool includes *AU* editor. In this *AU* editor, several kinds of expression can be generated by assigning the *Action Units'* numbers and their intensities in the window. The relation between AU parameters and human's emotion condition is expressed in *Emotion Space* in this system. An extraction of emotional features in several media is future problem.

5. ANALYSIS OF EMOTIONAL FEATURE OF VOICE[9]

Emotional feature of voice have not been studied so well until now. Even if the target of classification is limited only to typical emotional patterns, advanced media conversion from voice to face expression is possible. For example, when one is speaking with angry voice, face in the display gets angry. When one is speaking with sad voice, face appears to be sad. Emotion analysis and generation are important problem in speech synthesis too. In this section, we try to do a quantative analysis of voice including only five basic emotions.

5.1. Voice Sample with Emotion

Real actors speak some sentences and words with basic emotions and we recorded them as voice samples. These sentences and words don't include emotional meaning. Basic emotions are anger, sadness, happiness, disgust and neutral. Typical samples which are considered as representing basic emotions are selected by subjective test.

5.2. Voice Feature of Emotion

For the first trial, average power spectrum of octave band and histogram of pitch period are analized with emotional voice samples. Figure 2 indicates average power spectrums in octave band comparing with the power in 125 to 250[Hz] band. In both cases, there is an order of emotion according to the strength of high frequency power. This order appears to be *anger, happiness, disgust, neutral, sadness*. Figure 3 shows histograms of pitch period in voice. In both cases, there is the order of peaks. *Sadness* is longer than neutral, *Disgust* is shorter and *happiness/anger* are much shorter. However, the order of *happiness and anger* can not be decided by this result.

Now we are trying to analize all of the other acoustical feature parameters of sample voices too and it's our goal to find speaker- and sentence-independent parameters of emotional features.

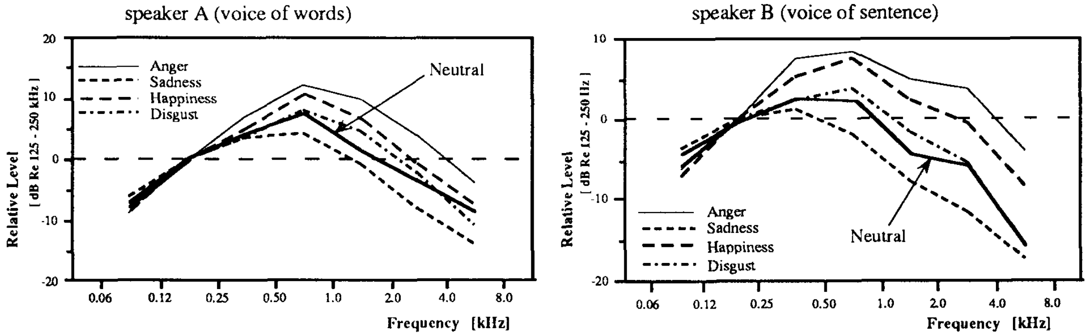

Figure 2. Average Power Spectrums in Octave Band

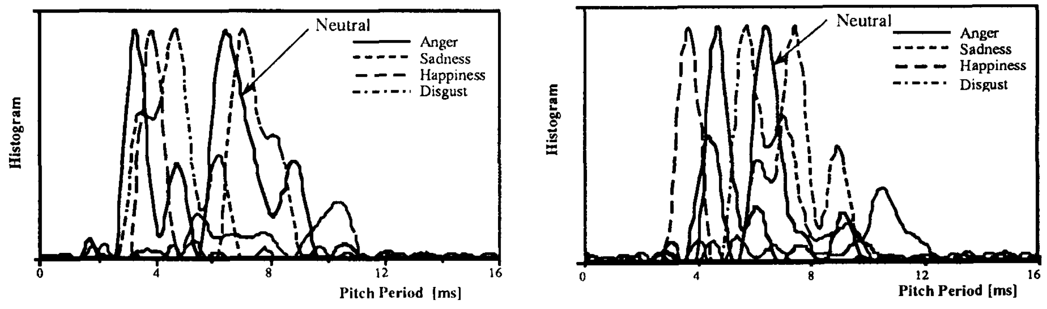

Figure 3. Histograms of Pitch Period in a Voice Samples

6. CONCLUSION

This paper presents a prototype of multi-media interface with face based on several kinds of media conversion schemes. Mouth shape and its motion can be synthesized by analysis of voice and sentence naturally. Scenario making tool gives the user friendly expression synthesis environment. In future, expression control based on the emotional feature in voice will be possible. For the first trial, some experimental results of voice feature analysis with basic emotion are presented. This technique will contribute to the advanced voice communication system and user-friendly human-machine interface.

References

[1] P.Ekman and W.Friesen, "Facial Action Coding System", Consulting Psychologists Press, 1977.

[2] S.Morishima and H.Harashima, "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface", IEEE Journal on Selected Areas in Communication, Vol.9, No.4, 1991.

[3] S.Morishima and H.Harashima, "Speech-to-Image Media Conversion Based on VQ and Neural Network", Proceedings of ICASSP91, M10.11, pp.2865-2868, 1991.

[4] S.Kobayashi, S.Morishima et.al, "Representation of Feel and Motion of the Thread-like Objects", Proc. of NICOGRAPH90, pp.29-36, 1990.

[5] Y.Fukuda and S.Hiki, "Characteristic of the mouth shape in the production of Japanese - Stroboscopic Observation", Journal of Acoustical Society of Japan (E), 3.2, pp.75-91, 1982.

[6] M.Potmesil and E.M.Hoffert, "A Pixel Machine: A Parallel Image Computer", ACM Computer Graphics, vol.23, No.3, pp.69-78, 1989.

[7] S.Morishima and H.Harashima, "A Proposal of a Knowledge Based Isolated Word Recognition", Proc. ICASSP, 14.5, 1986.

[8] C.S.Choi, H.Harashima and T.Takebe, "Analysis and Synthesis of Facial Expressions in Knowledge Based Coding of Facial Image Sequences", ICASSP91, M9.7, pp.2737-2740, 1991.

[9] C.E.Williams and K.N.Steven,: "Emotions and Speech: Some Acoustical Correlates", JASA, 52(4), pp.1238-1250, 1972.

[10] S.Morishima and H.Harashima,"Human Machine Interface Using Media Conversion and Model-Based Coding Schemes", Visual Computing, CG International Series, Springer-Verlag, pp.95-105, 1992.

[11] S.Morishima, T.Sakaguchi, H.Harashima, "A Facial Image Synthesis System for Human-Machine Interface", IEEE International Workshop on Robot and Human Communication, pp.363-368, 1992.

[12] E. Ono, S.Morishima and H.Harashima, "A model based shade estimation and reproduction schemes for rotational face", Picture Coding Symposium '93, 2.2, 1993.

[13] S.Morishima, T.Sakaguchi and H.Harashima, "Face animation scenario making system for model based image synthesis", Picture Coding Symposium '93, 13.19, 1993.

[14] S.Morishima and H.Harashima, "Facial Animation Synthesis for Human-Machine Communication System", HCI International, 1993.